International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

Sentiment Analysis of News Event-based Social Network using Data Mining Technique

Hlaing Phyu Phyu Mon[a*], Thin Thin San[b]

[a,b]*Faculty of Information Science, University of Computer Studies (Meiktila)*

[a]*Email: hlaingppmon@gmail.com*

[b]*Email: Drhppmon.is@ucsmtla.edu.mm*

Abstract

The increasing popularity of social media takes the attention of the internet users across the word wide to discuss and share the events/things they are interested on social media blogs/sites. Consequently, an explosive increase of social media data spread on the web has been promoting the development of analysis of social media news depending on the news or events, the latest trend of the social big data. The sentiment analysis of news event becomes an important research area for many real-world applications, such as public opinion monitoring for government and news recommendation of news websites. In this paper, we perform sentiment analysis for news events based on posts, and comments of the users upon a news event. We use two data mining techniques namely naïve Bayesian and support vector machine to reveal what the polarity/meaning of the post is such as positive, negative or polarity. There are two main stages in performing this task called training and testing phases. The first phase uses the training datasets of the news event and the second phase use newly inputted data of the user to classify the polarity of the user news posts or comments. We then execute the experiments for each algorithm and then collect the experimental results and compare them with accuracy with known and unknown test data with different volumes of tweet transactions. According to the results, both of them can accurately reveal the opinions of the social network users.

Keywords: sentiment analysis; support vector machine; naïve Bayesian; news-based social networking posts.

----

* Corresponding author.

1. Introduction

Web information enriches plentiful, wide knowledgeable including explicit and implicit knowledge. How to handle web interests of the users and how to use Web for better business applications has become the major topic in today internet world. In Myanmar, social network is the most popular thing among internet users. Almost all internet users use social networks on daily basis. They mostly debate current trending of Myanmar such as politics, economic, education. Among social networking websites, Facebook is the most popular and almost 10 million users are using Facebook and make some traffic on it upon the latest topic of Myanmar. As increasing the number of social network users, their feelings and opinion expressed upon the network is increasing day after day. The social media data is considered as the instance of big data organized with 4V named as volume, variety, velocity and value. The volume means the average number of social network posts/comments regarding a news event, which is at least 100,000 number on daily basis for 10 million users according to statistics of social network sites. There are different variety that can contain different materials such as words, emoticons, figures, etc. The velocity means the data used in social network sites are very high dynamic, which means there are 500TB in Facebook in each single day. The values are the rich information embedded in social media websites which need to be handled for further analysis, such as managing the data, recommendation news and exploring the opinions of users interests and feelings [1-9]. Among social network things, news events are a significant component among network users. The news articles in Facebook reflect updated conditions or news of the world and they are reported by the users in every different way. The popular and valuable information are shared by different sources and they are instantly spread to the people across the world. In Myanmar, Facebook becomes a place to share valuable information and advertainments place for those who are finding the public’s attention. Therefore, browsing and discussing the news events has become part of daily lives for the people. Among lots of news event analysis, the challenging task is to analysis the opinions of the news/events discussions to discover how people feels about them [16-19]. The ability to discover the sentiment of a news event can be applied in different ways such as public monitoring system upon the posts/comments of network users; product feedback monitoring system and locating the services available for those who wanted. This paper classifies the news event based social network posts and comments into three categories namely positive, negative and neural. However, analyzing the human language written in pure English text, emoticons and usage of slang is quite complexing because it needs to consider and compute text multidimensional attributes and data. We first of all, data are pre-processed before using data mining techniques. After preprocessing, we use data mining techniques called support vector machine and naïve Bayesian classification. The remainder of the paper is organized as follows. The related work regarding sentiment analysis is performed in Section 2. Section 3 widely explains about the proposed system regarding sentiment analysis using proposed data mining algorithms. The experimental results are demonstrated in Section 4 and the paper is concluded in Section 5.

2. Related Work

Sentiment analysis is to deal with variety of input texts such as posts, comments, emoticon and slang. It is a kind of text classification. Pang and his colleagues [14] exploits machine learning techniques to analyze the sentiment of movie reviews, and they divided the film review into two categories called positive and negative.

They applied three machine learning methods in the experiments and results are shown using bag of words as features. Cui and his colleagues [15] executed experiments on opinion classification to prove that unigram was superior with no too many training corpus, and with the increasing number of training corpus, n-gram played a very important role. Dandan and his colleagues [1] proposed an innovative method to do the sentiment computing for news events. They built a word emotion association network (WEAN) to jointly express its semantic and emotions, computational algorithm is proposed to obtain the initial words emotion, stored in standard emotion thesaurus. With the word's emotion prestored in the thesaurus, they compute every sentence's sentiment. The graph based opinioned post detector (GOPD) is proposed by Yanzhang and his colleagues [10] describes the opinioned similarity between posts of Sina Weibo Chinese social network websites. Their GOPD utilizes three types f user interaction, that includes reposting, responding and referring to construct the opinioned similarity graph (OSG), in order to describe the opinioned similarity between posts of Sina Weibo web. Feng and his colleagues [21] proposed a novel recommendation method which incorporates information on common author relations between articles. The rationale based on their method is that researchers often search articles published by the same authors. Since not all researchers have such author-based search patterns, they introduce two features, that are defined based on information about pariwise articles with common author relations and frequently appeared authors, to determine target researchers for recommendation. Baoshan and his colleagues [5] adopted clustering and time impact factor matrix to monitor the degree of user interest drift in the class and more accurately predict an item's rating. They added a time impact factor to the original baseline estimates and use the linear regression to predict the user interest drift. To make prediction, they applied short-term, long-term and periodic effects to the time impact factor. They utilize time distribution of dataset to calculate time decay function. They mainly use traditional clustering algorithm, and then improve it with element numbers relatively average in each class. Their target is to achieve better prediction for user internet drift. Danushka and his colleagues [3] presented a method to overcome the problem in cross-domain sentiment classification. First, they created a sentiment sensitive distributional thesaurus using labeled data for the source domains and unlabeled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document sentiment labels in the context vectors used as the basis for measuring the similarity between the words. They then exploited the created thesaurus to expand feature vectors during train and test times in a binary classifier. They exclaimed that their proposed method significantly outperforms numerous baselines and returns results that are comparable with previously proposed products.

3. Sentiment analysis of social network related new events

The users express their opinions and feelings about a news event upon the social network sites, review sites, etc. Reviews on a wide variety of communities are available on the web. The analysis for the news events, which aims to obtain the polarity (positive, negative, and neutral) from news event posts in Facebook, consists of two main parts: data pre-processing and data classification [10-13]. The first pre-processing steps is to clean the data to get to be ready for classification process, which is the second part that will classify if the input (user posts/comments in Facebook) is positive, negative or neural opinion. In classification, we apply two popular classification methods called naïve Bayesian and support vector machine.

3.1. Data Preprocessing

Data pre-processing is the form of cleaning that involves removal of stop words, extracting most common words from text etc [20]. In this system, we concern the following.

1. Build dictionary of words from input transactions from training set.
2. Consider the most common 3000 words.
3. For each document in training set, create a frequency matrix (as shown in Figure 1) for these words in dictionary and corresponding labels (pos, neg and neu for positive, negative and neutral).
4. That frequency matrix is considered as the features of the classification models.

$$\begin{array}{r|ccccccc|l}
 & w1 & w2 & w3 & .. & .. & .. & wm & \\
\hline
doc1 & 3 & 4 & 5 & .. & \ldots. & .. & 7 & pos \\
doc2 & 2 & .. & & & & & 8 & neg \\
\cdots & 7 & \ldots. & & & & & 4 & .. \\
docn & 10 & .. & & & .. & & 12 & neu
\end{array}$$

Figure 1: Feature matrix of input tweets

3.2. Data Classification

In classification the sentiment text, we perform two popular classification methods called naïve Bayesian and support vector machine as described in following sections.

3.2.1. Naïve Bayesian Classification of sentimental data

Naïve Bayes (NB) classifiers are probabilistic classifiers, meaning that they use the probabilities of observed outcomes to return a reasonable estimate of an unknown outcome. At a higher level, NB classifiers use Bayes's rule with one naïve (or simplifying) assumption: that features are independent from/uncorrelated with other features. In our system, we use following Bayes' theory for calculation of the percentages of the polarity values.

$$\Pr(x|y) = \frac{\Pr(x)\Pr(y|x)}{\Pr(y)} \qquad (1)$$

In sentiment analysis, we might prefer using binary multinomial, which simply counts 1 if a word appears and 0 otherwise, since we really only care about a word appearing or not in, say, a positive review, rather than how many times it appeared. In eq (2), y is the class label for classifying the data x. With that probability values, we can compute a word found in various conditions (various polarity labels). We regard that the input sentence (sentiment) have what kind of polarity it has depending on probability values of each individual word contained in each sentence against with different values. For example, in the following calculation, we assume like will be more likely to be found together with positive class labels instead of matching with other twos.

$$P(like|pos) = \frac{\Pr(like)\Pr(pos|like)}{\Pr(pos)} \qquad (2)$$

$$P(like|neg) = \frac{\Pr(like)\Pr(neg|like)}{\Pr(neg)} \qquad (3)$$

$$P(like|neu) = \frac{\Pr(like)\Pr(neu|like)}{\Pr(neu)} \qquad (4)$$

3.2.2. Support Vector Machine of sentimental data

A support vector machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new example. The SVM algorithm is implemented in practice using a kernel, which is a trick to take low dimensional input space and transform it to a higher dimensional space. It converts non-separable problem to separable problem, used for both linear and non-linear problems. In this paper, we use SVM to group each data item as a point in n-dimensional space in which n is the number of features obtained from previous preprocessing steps. We then use Linear SVM for classification the input text features to produce their polarity values (pos, neg, neu, etc) using the following algebraic equations.

$$y = B(x) + sum(a_i * x_i) \qquad (5)$$

where B and a are the tuning parameter chosen by SVM for input data attributes (x_i) to guess the labels of y.

4. Experimental results

In order to compare the efficiencies of two applied classification algorithms, we conduct some experiments with Tweets dataset which is publicly available from UCI machine learning repository [22]. We split 2/3 of this dataset as training data and the rest is as testing phase. Besides, the newly added sentiments are applied for unknown testing data. Results are collected and illustrated as follows. Accuracy measurement is performed to check the overall accuracy of the system with the ratio of number of correct classification results to the total number of classifications. In comparison of two algorithms, we found that SVM better performs than NB when the input features contains many dimensions (attributes). In low dimensional data, both of them can execute similar classification results, which is quite acceptable for both training and testing data. As shown in figure 2, the more dimension they have, they more accurate results they can reveal, especially for SVM. SVM works better with higher dimensions and is more robust with incomplete data. To evaluate the classification quality, we feed different data called known (part of dataset) and unknown testing data and extract the results as shown in Figure 3. Intuitively, known data can achieve more accurate results in NB while SVM brings better results than NB in unknown test data.

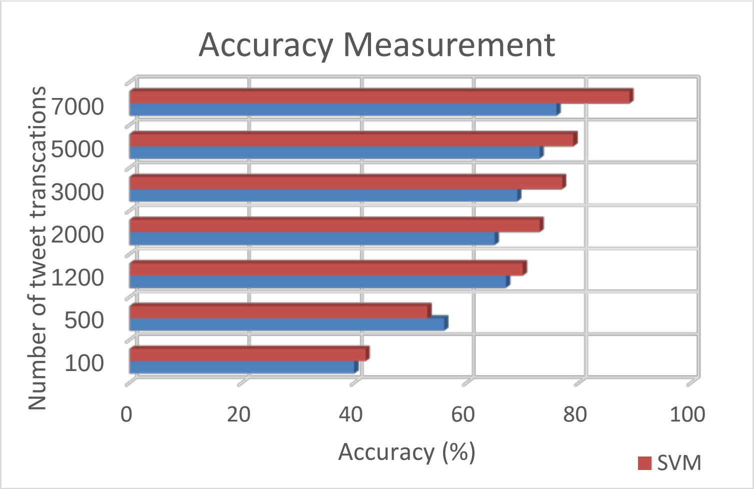

Figure 2: Accuracy Measurement of NB and SVM algorithms

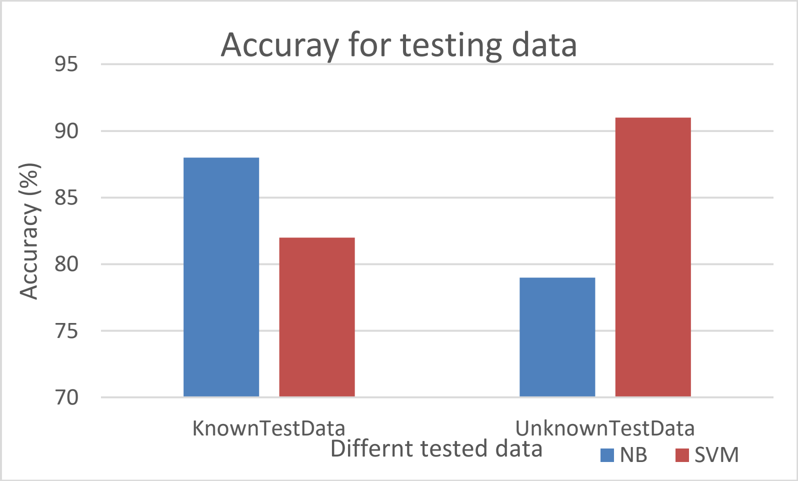

Figure 3: Accuracy Measurement of different tested data

5. Conclusions

With the proliferation of social media over the last decade, the demand of analyzing the people's attitudes with respect to a specific topic, document, interaction and events, has been increasing in order to know the interest trends of the people so that business or governmental benefits can be taken. In this paper, we implement two classification algorithms called naïve Bayesian and support vector machine to analysis the polarity values of tweets sentence from social network users. We afterwards evaluate classification results of both of those algorithms so as to reveal which algorism does work well in what kind of situations such as known, unknown data with different volumes of tweet transactions. We will extend this work by adding required factors and conditions in order to achieve better classification results.

References

[1]. Dandan Jiang, Xiangfeng Luo, Junyu Xuan, Zheng Xu, Sentiment Computing for the News Event

based on the social media big data, Special Section on intelligent sensing on mobile and social media analytics, IEEE Access, vol. 5, 2017, pp-2373-2382.

[2]. Xiaohui Yu, Yang Liu, Jimmy Xingji Huag, Aijun An, Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain, IEEE transactions on knowledge and data engineering, vol 24, no 4, 2012, pp-720-734.

[3]. Danushka Bollegal, David Weir, John Carroll, Cross-Domain Sentiment Classification using a sentiment sensitive thesaurus, IEEE transactions on knowledge and data engineering, Vol 25, No. 8, 2013, pp-1719-1731.

[4]. Zheng Xu, Shunxiag Zhang, Kim-kwang Raymond Choo, Lin Mei, Xiao Wei, Xiangfeng Luo, Chuanping, Yunhuai Liu, Hierarchy-cutting model based association semantic for analyzing domain topic on the web, IEEE transactions on industrial informatics, vol. 13, no. 4, 2017, pp-1941-1950.

[5]. Baoshan Sun and Lingyu Dong, Dynamic Model Adaptive to User internet drift based on cluster and nearest neighbors, IEEE Access , vol 5, 2017, pp-1682-1691.

[6]. X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach,” in Proc. 20th ACM Int. Conf. Inform. Knowl. Manag., pp. 1031–1040, 2011.

[7]. L. Zhuang, F. Jing, and X. Zhu, “Movie review mining and summarization,” in Proc. ACM 15th Conf. Inform. Knowl. Manag., Arlington, Virginia, USA, Nov. 2006, pp. 43–50.

[8]. G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” J. Artif. Intell. Res., vol. 22, pp. 457–479. 2004.

[9]. R. McDonald, “A study of global inference algorithms in multidocument summarization,” in Proc. 29th Eur. Conf. IR Res., 2007, pp. 557–564.

[10]. Yanzhang Lv, Jung Liu, Hao Chen, Jianhong Mi, Mengyue Liu, Ainghua Zheng, Opinioned Post Detection in Sian Weibo, Special Section on trust management in pervasive social networking, IEEE Access, 2017, pp-7263-7271.

[11]. S. Zhang, J. Liang, R. He, and Z. Sun, ``Code consistent hashing based on information-theoretic criterion," IEEE Trans. Big Data, vol. 1, no. 3, pp. 84_94, Sep. 2015.

[12]. J. Xuan et al., ``Uncertainty analysis for the keyword system of web events," IEEE Trans. Syst., Man, Cybern., Syst., vol. 46, no. 6, pp. 829_842, 2016.

[13]. W. McDougall, An Introduction to Social Psychology. North Chelmsford, MA, USA: Courier Corporation, 2003.

[14]. B. Pang, L. Lee, and S. Vaithyanathan, ``Thumbs up?: Sentiment classification using machine learning techniques," in Proc. ACL Conf. Empirical Methods Natural Lang. Process., 2002, pp. 79_86.

[15]. H. Cui, V. Mittal, and M. Datar, ``Comparative experiments on sentiment classification for online product reviews," in Proc. AAAI, vol. 6. 2006, pp. 1265_1270.

[16]. S.-M. Kim and E. Hovy, ``Automatic identification of pro and con reasons in online reviews," in Proc. COLING/ACL Main Conf. Poster Sessions, 2006, pp. 483_490.

[17]. Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth, Challenges of Sentiment Analysis for Dynamic Events, IEEE intelligent system, 2017, pp-70-75.

[18]. Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized travel sequence recommendation Multi-

Source big social media, IEEE transaction on big data, vol. 2, no. 1, 2016, pp-43-56.

[19]. H. Liu, T. Mei, J. Luo, H. Li, and S. Li, “Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing,” in Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 9–18.

[20]. J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, “GPS estimation for places of interest from social users’ uploaded photos,” IEEE Trans. Multimedia, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[21]. S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, “Author topic model based collaborative filtering for personalized POI recommendation,” IEEE Trans. Multimedia, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[22]. J. Sang, T. Mei, T. J. Sun, S. Li, and C. Xu, “Probabilistic sequential POIs recommendation via check-in data,” in Proc. ACM SIGSPATIAL Int. Conf. Adv. Geographic Inform. Syst., 2012, pp. 402–405.

[23]. Feng Xia, Haifeng Liu, Longbing Cao, Scientific article recommendation: exploiting common author relations and historical preferences, IEEE transactions on big data, vol. 2, no.2, 2016, pp-101-112.

[24]. https://github.com/dkakkar/Twitter-Sentiment-Classifier